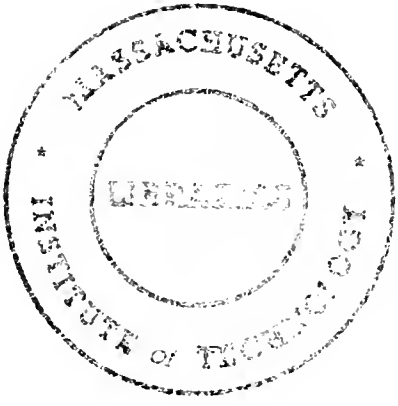

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY

A kth Nearest Neighbour Clustering Procedure

M. Anthony Wong and Tom Lane

WP#1213.81

May 1981

A kth Nearest Neighbour Clustering Procedure

M. Anthony Wong and Tom Lane

WP#1213.81 May 1981

A kTH NEAREST NEIGHBOUR CLUSTERING PROCEDURE

M. Anthony Wong and Tom Lane

Sloan School of Management

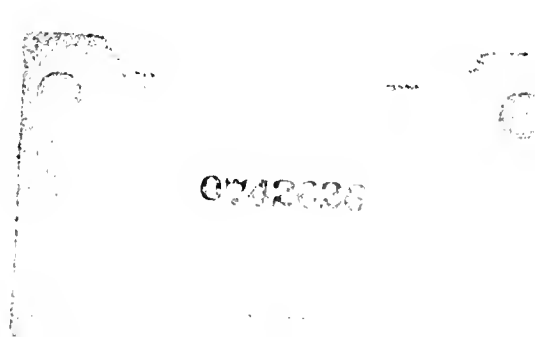
Massachusetts Institute of Technology

Cambridge, MA 02139

SUMMARY

Due to the lack of development in the probabilistic and statistical aspects of clustering research, clustering procedures are often regarded as heuristics generating artificial clusters from a given set of sample data. In this paper, a clustering procedure that is useful for drawing statistical inference about the underlying population from a random sample is developed. It is based on the uniformly consistent kth nearest neighbour density estimate, and is applicable to both case-by-variable data matrices and case-by-case dissimilarity matrices. The proposed clustering procedure is shown to be asymptotically consistent for high-density clusters in several dimensions, and its small-sample behavior is illustrated by empirical examples. A real application is also included to demonstrate the practical utility of this clustering method.

Keywords: CLUSTERING PROCEDURE; HIGH-DENSITY CLUSTERS; kTH NEAREST NEIGHBOUR DENSITY ESTIMATION; SET-CONSISTENCY.



1. INTRODUCTION

1.1 Shortcomings of Clustering Procedures

A recent study by Blashfield and Aldenderfer (1978) shows that numerous clustering methods have been developed in the past two decades. A review of many of these techniques can be found in Cormack (1971), Anderberg (1973), Sneath and Sokal (1973), Everitt (1974), Hartigan (1975), and Spath (1980). However, hardly any of the originators of these methods have approached the clustering problem from within a theoretical framework. More often than not, the concept of a real population cluster is vague and is left undefined. Since no statistical evaluation of the sample clusters can be performed under the circumstance, the validity of the clusters obtained by these methods is always questionable. Consequently, the existing clustering procedures are often regarded as heuristics generating artificial clusters from a given set of sample data, and there is a need of clustering procedures that are useful for drawing statistical inference about the underlying population from a sample. In this paper, a clustering procedure based on the kth nearest neighbour density estimate is proposed, and it is shown to be set-consistent for high-density clusters in several dimensions. The set-consistency property of a hierarchical clustering procedure will be defined next.

1.2 A Theoretical Approach to Evaluating Hierarchical Clustering Methods

In order to evaluate the sampling property of a clustering method, it is necessary to have population clusters defined on population probability density functions from which the observations are obtained, and to have some ways of judging how the sample clusters deviate from the population clusters. Let the observations $x_1, x_2, \ldots, x_N$ in p-dimensional space be sampled from a population with density f, taken with respect to Lebesque measure. Using the

high-density clustering model given in Hartigan (1975, p. 205), the true population clusters can be defined on f as follows: a high-density cluster at level f_o in the population is defined as a maximal connected set of the form $\{x|f(x) \geq f_o\}$. The family T of such clusters forms a tree, in that $A \in T$, $B \in T$ implies either $A \supset B$, $B \supset A$, or $A \cap B = \phi$. A hierarchical clustering procedure, which produces a sample clustering tree T_N on the observations x_1, ..., x_N, may then be evaluated by examining whether T_N converges to T with probability one when N approaches infinity. A clustering method (or equivalently, T_N) is said to be strongly set-consistent for high-density clusters (or T) if for any A, $B \in T$, $A \cap B = \phi$,

$$P_r\{A_N \cap B_N = \phi \text{ as } N \to \infty\} = 1,$$

where A_N and B_N are respectively the smallest cluster in the sample tree T_N containing all the sample points in A and B. Since $A \subset B$ implies $A_N \subset B_N$, this limit result means that the tree relationship in T_N converges strongly to the tree relationship in T.

Using this definition of consistency, hierarchical clustering methods can be evaluated by examining whether they are strongly set-consistent for high-density clusters. If a clustering procedure is set-consistent, the sequence of enlarging hierarchical clusters that it produces in the sample are groups of points lying within successively lower density contours in the underlying distribution. Hence, these sample high-density clusters are useful in indicating the number of modal regions in the population as well as identifying their locations in the underlying space. And since it is the geometrical shape of the population density contours that determines the configuration of the sample high-density clusters, a set-consistent clustering method does not impose structure on the clusters it produces. (See Everitt (1974, Chapter 4) for some well-known clustering methods that impose a

spherical structure on the clusters they produce.) On the other hand, a clustering procedure that is not set-consistent is not adaptive to the underlying distribution, and is hence not suitable for detecting high-density or "natural" clusters (see Carmichael et. al., 1968).

Hartigan (1977a, 1977b, 1979) has examined the set-consistency of many of the best known hierarchical clustering methods for high-density clusters. It was shown that the complete linkage (Sorenson 1948) and average linkage (Sneath and Sokal 1973) methods are not set-consistent, while single linkage (Sneath 1957) is weakly set-consistent in one dimension but not in higher dimensions. Thus most of the relevant evaluative work under the high-density clustering model have been carried out. However, the important problem of developing clustering procedures that are set-consistent for high-density clusters did not receive much attention. In Hartigan and Wong (1979), and Wong (1980), a hybrid clustering method is developed which is weakly set-consistent for high-density clusters in one dimension; and, there exist empirical evidence that similar consistency results hold in several dimensions. However, although the hybrid method has the advantage of being practicable for very large data sets, it is not well-suited for small samples ($n < 100$) and it is only applicable to case-by-variable data matrices. In this paper, a strongly set-consistent clustering procedure is developed which is applicable to both case-by-variable data matrices and case-by-case distance matrices, and its development is outlined next.

1.3 Development of the k_nth Nearest Neighbour Clustering Procedure

Under the high-density clustering model, density estimates can be used to generate sample clusters, namely the high-density clusters defined on the estimates. And a clustering procedure is expected to be set-consistent for high-density clusters if it is based on a uniformly consistent density

estimate. Single linkage corresponds to nearest neighbour density estimation (Hartigan 1977b), in which the density estimate $f_N(x)$ at a point x is inversely proportional to the volume of the smallest closed sphere including one sample point. This density estimate is not consistent in the sense that $f_N(x)$ does not approach f(x) in probability. An improved density estimate, and perhaps improved clustering, can be obtained by the kth nearest neighbour density estimate: the estimated density at point x is $f_N(x) = k/(N\ V_k(x))$, where $V_k(x)$ is the volume of the closed sphere centered at x containing k sample points. Such a density estimate is uniformly consistent with probability 1 if f is uniformly continuous and if $k = k(N)$ satisfies $k(N)/N \to 0$ and $k(N)/\log N \to \infty$. (See, for example, Devroye and Wagner 1977, and Moore and Yackel 1977.)

Wishart (1969), in an attempt to improve on the single linkage clustering technique, developed a procedure entitled Mode Analysis which is related to the kth nearest neighbour density estimate. However, Wishart's procedure was not designed to obtain the high-density clusters defined on the density estimate, and hence its set-consistency for high-density clusters was never established. Moreover, since its computational algorithm is quite complicated, the Mode Algorithm did not receive much attention in the clustering literature. In this paper, a clustering algorithm for deriving the tree of sample high-density clusters from the kth nearest neighbour density estimate is developed. A detailed description of this clustering procedure is given in Section 2. In Section 3, it is established that the proposed method is strongly set-consistent for high-density clusters. Empirical examples are given in Section 4 to illustrate the small-sample behavior of kth nearest neighbour clustering. A real example is presented in Section 5 to demonstrate the practical utility of the proposed clustering method.

2. A KTH NEAREST NEIGHBOUR CLUSTERING PROCEDURE

The proposed nearest neighbour clustering algorithm consists of two stages. At the first stage, the kth nearest neighbour density estimation procedure is used to obtain a uniformly consistent estimate of the underlying density. The tree of sample high-density clusters defined on the estimated density is computed at the second stage of the algorithm. At this latter stage, a distance matrix is first computed in which the distance between two "neighbouring" points (i.e. points with the property that at least one point is one of the kth nearest neighbour of the other) is defined to be inversely proportional to a pooled density estimate at the point halfway between them, and the single linkage clustering algorithm (Sneath, 1957) is then applied to this distance matrix to obtain the tree of sample clusters.

2.1 The Density Estimation Stage

The kth nearest neighbour density estimation procedure is used in this stage of the clustering procedure because it provides a strongly uniform consistent estimate of the underlying density. Let $x_1,\ldots,x_N$ be independent, identically distributed random vectors with values in R^p, $p \geq 1$, and with a common probability density f. If $V_k(x)$ is the volume of the smallest sphere centered at x and containing at least k of the random vectors $x_1,\ldots,x_N$, then the kth nearest neighbour density estimate of f at x is

$$f_N(x) = k/(NV_k(x))$$

And in Devroye and Wagner (1977), the following strong uniform consistency result of this estimate is shown:

Lemma (Devroye and Wagner, 1977):

If f is uniformly continuous on R^p and if $k = k(N)$ is a sequence of positive integers satisfying:

(a) $k(N)/N \to 0$, and

(b) $k(N)/\log N \to \infty$, as $N \to \infty$,

then

$$\sup_x |f_N(x) - f(x)| \to 0 \text{ with probability 1.}$$

One purpose of the kth nearest neighbour clustering method is to discover the population high-density clusters given a random sample from some underlying distribution F with density f. In this first step of the proposed procedure, a uniformly consistent estimate of f is obtained. The high-density clusters defined on the estimated density f_N can then be used as sample estimates of the population high-density clusters defined on f. These hierarchical sample high-density clusters are constructed in the second stage of the proposed clustering algorithm.

2.2 The Hierarchical Clustering Stage

In this stage, a distance matrix $D(x_i,x_j)$, $1 \le i, j \le N$, for the N observations is first computed using the following definitions:

Definition 1: Two observations x_i and x_j are said to be neighbours if $d^*(x_i,x_j) \le d_k(x_i)$ or $d_k(x_j)$, where d^* is the Euclidean metric and $d_k(x_i)$ is the kth nearest neighbour distance to point x_i.

Definition 2: The distance $D(\cdot,\cdot)$ between the observations x_i and x_j is

$$D(x_i,x_j) = (1/2)[1/f_N(x_i) + 1/f_N(x_j)] = \frac{N}{2k}[V_k(x_i) + V_k(x_j)], \text{ if } x_i \text{ and } x_j \text{ are neighbors;}$$

$$= \infty, \text{ otherwise.}$$

Hence, finite distances are defined only for pairs of observations which are in the same neighbourhood in R^p, and the defined distance between a pair of neighbouring observations is inversely proportional to a pooled density estimate at the point halfway between them. The following single linkage

clustering technique is then applied to this distance matrix D to obtain the tree of sample high-density clusters.

Given a set of observations of objects $x_1, \ldots, x_N$ with distances $D(x_i,x_j)$, $1 \leq i < j \leq N$, single linkage clusters are defined as follows: let x_i and x_j be the closest pair of objects; amalgamate them to form a cluster c and define the distance between that cluster and any object x_z be $D(c,x_z) = \min [D(x_i,x_z),D(x_j,x_z)]$; repeat the process treating c as an object and ignoring x_i and x_j. The amalgamation continues until all objects are grouped in one large cluster. All clusters obtained in the course of this hierarchical algorithm are single linkage clusters. (See Gower and Ross (1969), and Hartigan (1975) for computational single linkage algorithms.) Single linkage clustering is used in this step of the proposed procedure because it has the following property: at every stage of the clustering, the single linkage clusters are the maximal linked sets if objects x_i and x_j are said to be linked whenever $D(x_i,x_j)$ is no greater than a given distance D_o. Now, since the distance D between two "neighboring" observations is reciprocal to the density estimate f_N at the midpoint between them, every cluster obtained by applying single linkage to D has the property that the density estimates over the objects in this cluster are greater than a certain density level f_o. Moreover, as the distance measure D is defined only for pairs of "neighbouring" observatons, the resultant single linkage clusters correspond to maximal connected sets of the form $\{x | f_N(x) \geq f_o\}$, which are the high-density clusters defined on f_N.

2.3 The Computational Algorithm

Since high-density clusters are invariant to monotone transformations of the density function, the kth nearest neighbour distances $d_k(x_i)$, $i = 1, \ldots, N$ are used instead of the $V_k(x_i)$'s in the following computational algorithm of

the kth nearest neighbour clustering procedure:

STEP 1: For i = 1, 2, ..., N, compute $d_k(x_i)$, the kth nearest neighbour distance of x_i. (For a computationally efficient algorithm to find the kth nearest neighbour distances, see Friedman et. al., 1975.)

STEP 2: Compute the distance matrix D as follows:

$D(x_i,x_j) = (1/2)[d_k(x_i) + d_k(x_j)]$ if

$d^*(x_i,x_j) \leq d_k(x_i)$ or $d_k(x_j)$, where d^* is the Euclidean metric;

$= \infty$, otherwise.

STEP 3: Apply the single linkage clustering algorithm to the computed distance matrix D to obtain the sample tree of high-density clusters.

The computational requirements for STEP 1 and STEP 3 are O(p N log N) and O(nk) respectively. Hence, unlike the hybrid clustering method (Hartigan and Wong, 1979 and Wong, 1980), this procedure is not practicable for large data sets; but, it is better suited for small samples, and is applicable to both case-by-variable data matrices and case-by-case dissimilarity matrices.

3. STRONG SET-CONSISTENCY OF kTH NEAREST NEIGHBOUR CLUSTERING

The asymptotic consistency of the kth nearest neighbour clustering method for high-density clusters in R^p, $p \geq 1$, is given in the following theorem:

Theorem: Let f denote a positive, uniformly continuous function on R^p such that $\{x | f(x) \geq f_o\}$ is the union of a finite number of compact subsets of R^p for every $f_o > 0$. Let T be the tree of population high-density clusters defined on f. Supose that A and B are any two disjoint high-density clusters in T with connected interiors. Let $x_1, \ldots, x_N$ be a random sample from f and let T be the hierarchical clustering specified by the kth nearest neighbour

clustering algorithm. Then, provided that $k = k(N)$ satisfies

(a) $k(N)/N \to 0$, and

(b) $k(N)/\log N \to \infty$.

as $N \to \infty$, there exist A_N, $B_N \in T_N$ with $A_N \supset A_N \cap \{x_1,\ldots,x_N\}$, $B_N \supset B \cap \{x_1,\ldots,x_N\}$ and $A_N \cap B_N = \phi$ with probability 1.

Proof: Since T_N is the tree of high-density clusters for f_N, this theorem is a direct consequence of the Lemma, which states that

$$\sup_x |f_N(x) - f(x)| \to 0, \text{ with probability 1.} \tag{3.1}$$

By definition, for any two disjoint high-density clusters A and B in T, there exist $\delta > 0$, $\epsilon > 0$ and $\lambda > 0$, such that

(i) $f(x) \geq \lambda$ for all $x \in A \cup B$, and (3.2)

(ii) each rectilinear path between A and B contains a segment, with length greater than δ, along which the density $f(x) < \lambda - 3\epsilon$. (3.3)

From (3.1), we have for N large,

$\sup_x |f_N(x) - f(x)| < \epsilon$ w.p. 1.

Thus, it follows fromm (3.2) and (3.3) that for N large, with probability 1,

(iii) $f_N(x) > \lambda - \epsilon$ for all $x \in A \cup B$, and

(iv) each rectilinear path between A and B contains a segment, with length greater than δ, along which the density estimate $f_N(x) < \lambda - 2\epsilon$.

Since A and B are disjoint, it follows from (3.4) and (3.5) that high-density clusters of the form $\{x | f_N(x) \geq \lambda - \epsilon\}$ separate the observations in A and B. The theorem follows.

4. EMPIRICAL STUDY OF THE SMALL-SAMPLE BEHAVIOR OF THE kTH NEAREST NEIGHBOUR CLUSTERING PROCEDURE

To illustrate the small-sample behavior of the kth nearest neighbor clustering procedure, an empirical study was performed in which the procedure is applied to various generated data sets. Results of three experiments, in which bivariate data were used, are reported here.

1. Experiment One: 30 observations were generated so that two spherical clusters of observations are present in this data set. The scatter-plot of this sample set is shown in Figure 1a, in which the observation numbers are plotted next to the observations. This data set is useful for illustrating the effectiveness of the proposed procedure in identifying spherical clusters. The dendrogram giving the hierarchical clustering obtained by the kth nearest neighbour method (using k = 4) is shown in Figure 1b. It is clear that, in this experiment, the kth nearest neighbour clustering indicates the presence of two modal regions or clusters.

However, for a choice of k which is much too small, the dendrogram produced by the hybrid method would tend to suggest the presence of a few extra modal regions. The reason for this moderate sensitivity of the proposed method to the choice of k is that, if k is too small, extra modes tend to appear in the kth nearest neighbour density estimate, and these bumps in the estimated density function are identified as modal regions in the hierarchical clustering stage of the algorithm.

2. Experiment Two: 58 observations were generated so that two elongated, elliptical clusters of observations are present in this data set. The scatter plot of this sample set is shown in Figure 2a, in which the observation numbers are plotted next to the observations. This data set is useful for

illustrating the effectiveness of the proposed clustering procedure inidentifying non-spherical clusters. The dendrogram giving the hierarchical clustering obtained by the kth nearest neighbour method (using k = 4) is shown in Figure 2b. Two disjoint modal regions, corresponding to the two elliptical clusters of observations shown in Figure 2a can be identified in this dendrogram. However, observations 51, 58, and 37 form a minor modal region within one of the two clusters; and, observations 22, 23, and 2 form a minor modal region in the other cluster.

3. Experiment Three: 60 observations were generated so that two spherical clusters of observations are present in this sample and they are connected by a chain of "noise" observations (see Figure 3a). This data set is useful for demonstrating the effectiveness of the proposed method when a moderate amount of noise is present in the sample. The hierarchical clustering obtained by the kth nearest neighbour method (using k = 4) is shown in Figure 3b. It can be seen that the two spherical clusters are recovered by the proposed method as modal regions, in spite of the presence of the noise observations.

5. A REAL EXAMPLE

In order to illustrate how the hybrid clustering method works in practice, it is applied to the well-known Iris data given in Fisher (1936). The data consist of four characteristics for three species of Iris; the species are Iris Setosa, Iris Versicolor and Iris Virginica, and the characteristics are sepal length, sepal width, petal length, and petal width. There are fifty samples from each species, and hence the total sample size is 150. This data set has been used by many authors to test the practical utility of various clustering algorithms (e.g. Friedman and Rubin, 1967). It has been found that there are two distinct clusters of samples in this data

set; one corresponding to the samples from Iris Setosa, and the other corresponding to samples from the other two species. Moreover, the samples from Iris Versicolor and Iris Virginica form a somewhat homogeneous group and there is no clearcut distinction between samples from these two different species (see, for example, Fisher, 1936; Friedman and Rubin, 1967; and Gnanadesikan, 1977).

The hierarchical clustering obtained by applying the kth nearest neighbour method is this data set, using a value of k = 8, is shown in Figure 4. Two distinct modal regions can be identified in this sample tree of high-density clusters; one corresponding to the samples from Iris Setosa, and the other corresponding to the samples from Iris Versicolor and Iris Virginica. Moreover, within the Versicolor-Virginica modal region, there are two sub-modal regions, one such region is consisted of samples only from Iris Versicolor, while the other region is consisted of samples only from Iris Virginica. However, it should be pointed out that if k was chosen to be much longer than 8 (say, k = 12 or 15), the two sub-modal regions would not appear in the hierarchical clustering and only the two well-known distinct clusters can be identified by the kth nearest neighbour method.

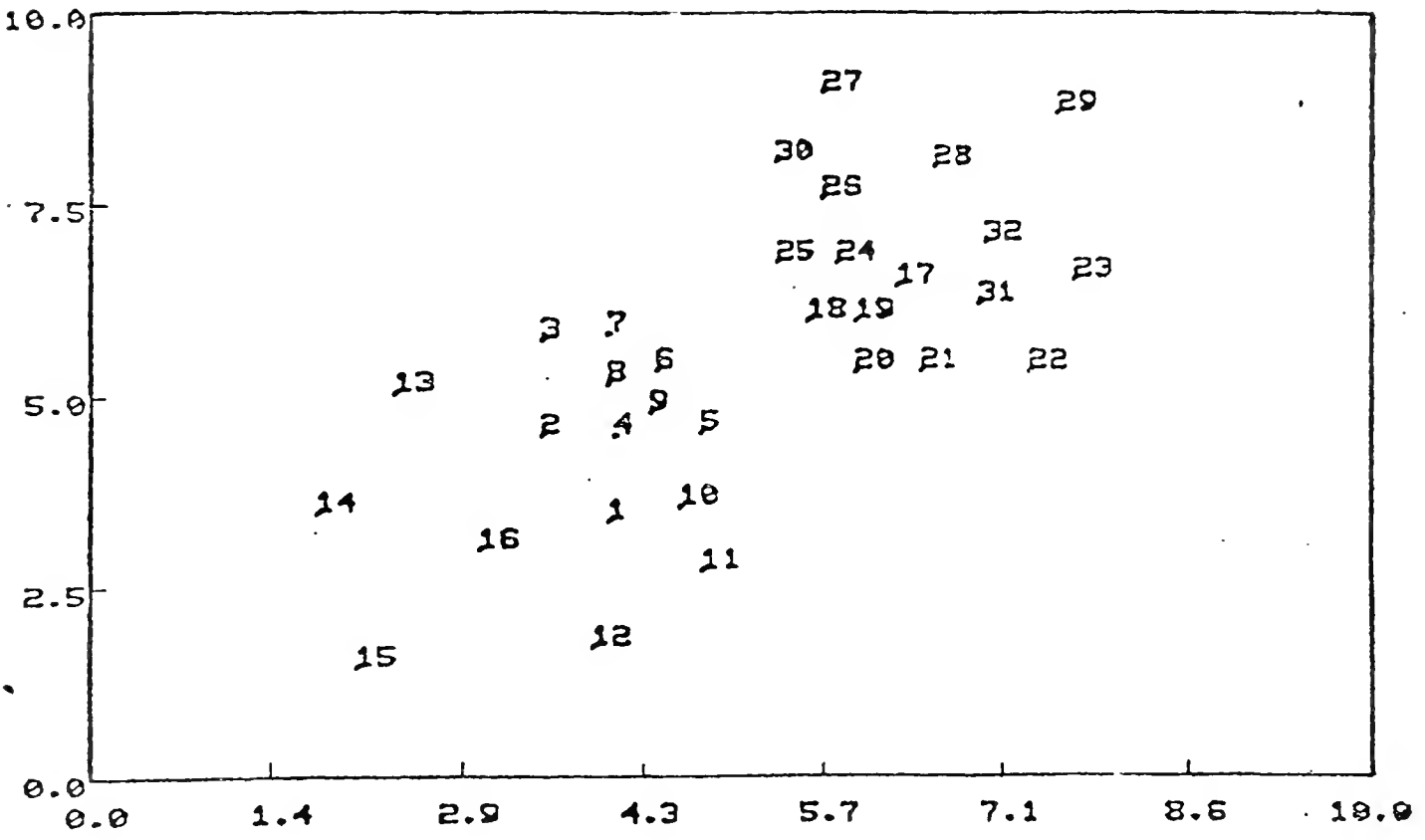


Fig. 1a. Scatter-plot of the generated bivariate sample (N = 30) used in Experiment One. Observation numbers are plotted next to the observations.

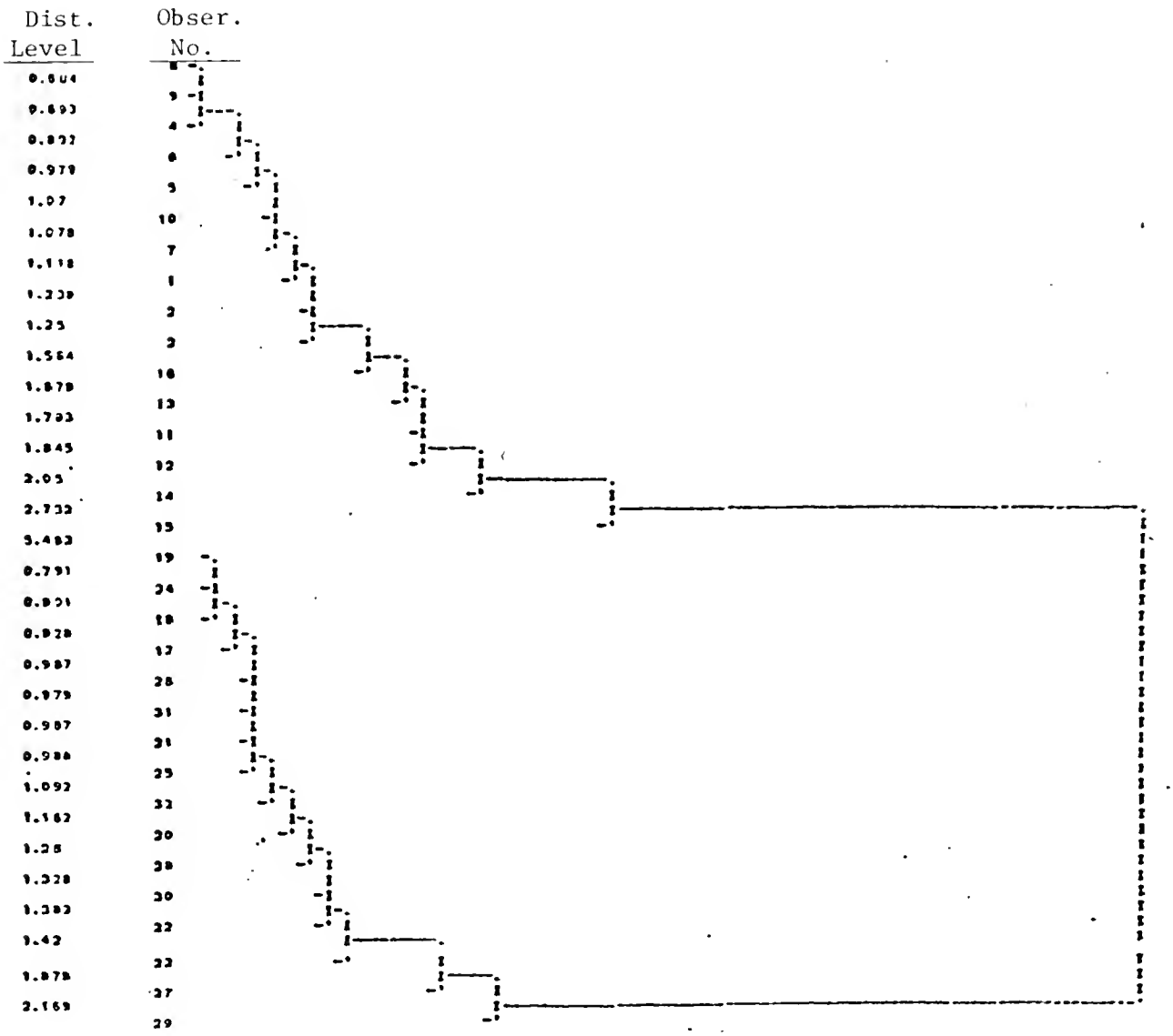


Fig. 1b. Tree of sample high-density clusters for data shown in Fig. 1a, derived from the kth nearest neighbour density estimate using k = 4.

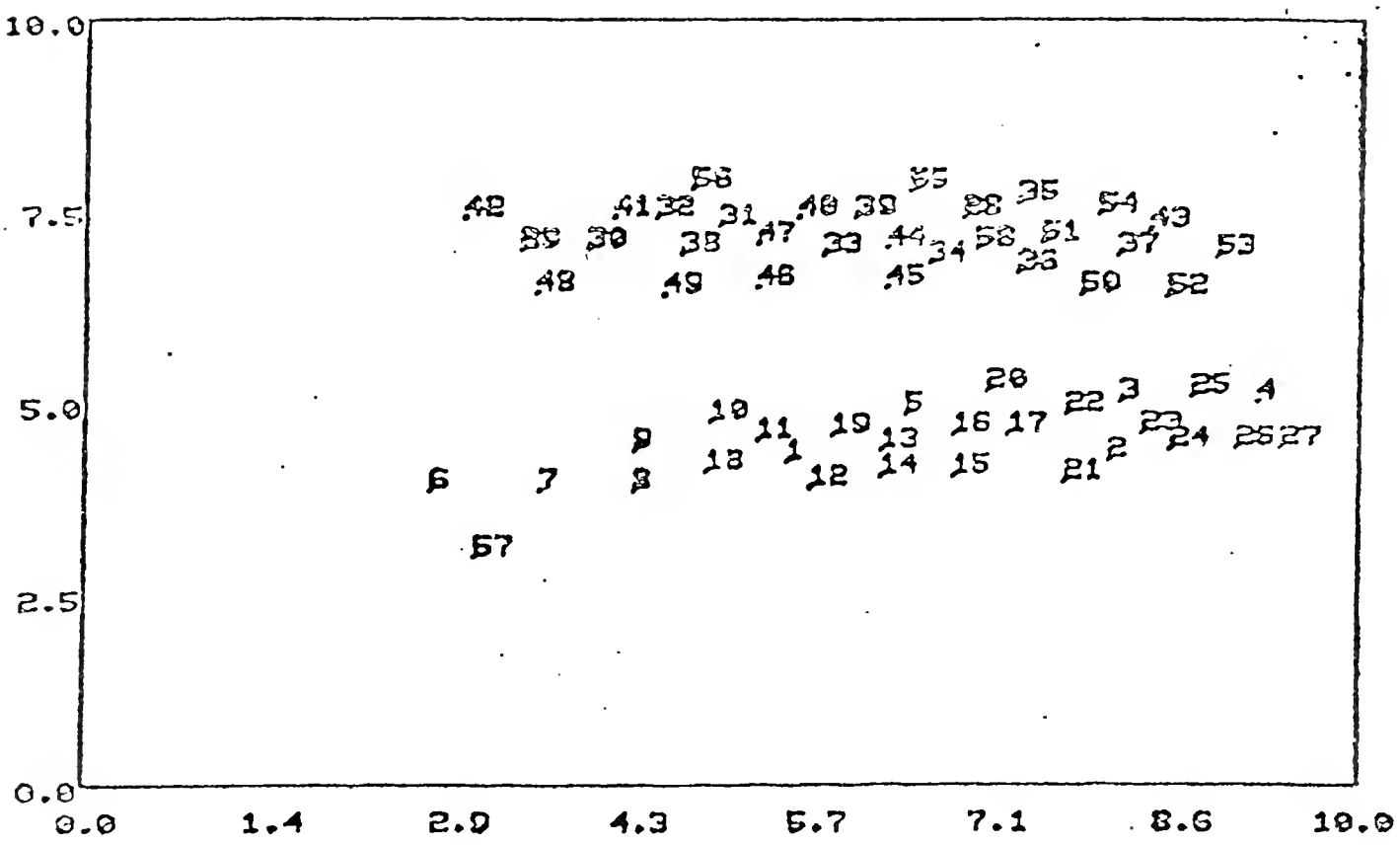


Fig. 2 a Scatter-plot of the generated bivariate sample (N = 58) used in Experiment Two. Observation numbers are plotted next to the observations.

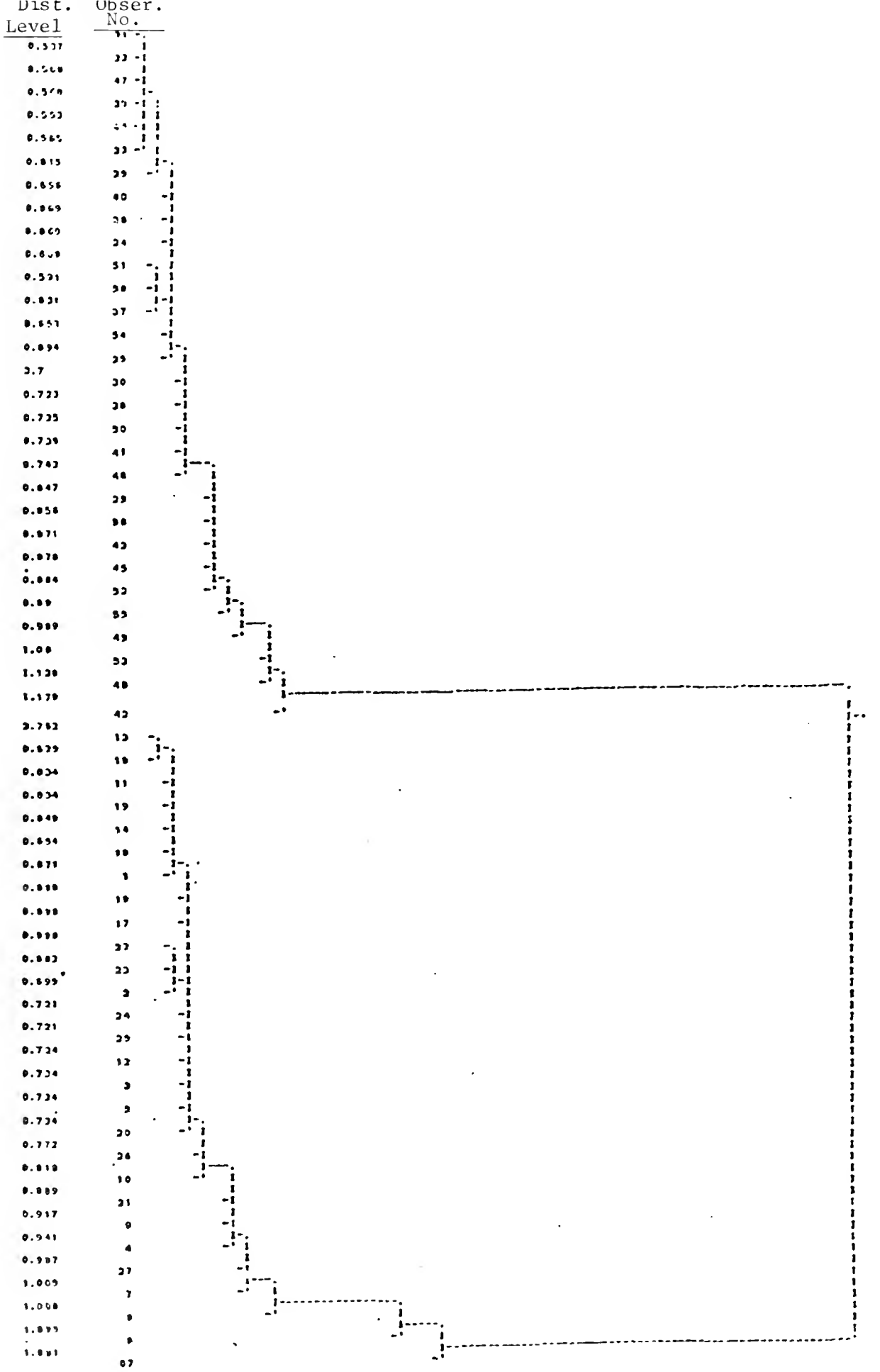


Fig. 2b Tree of sample high-density clusters for data shown in Figure 2a, derived from the kth nearest neighbour density estimate using k = 4.

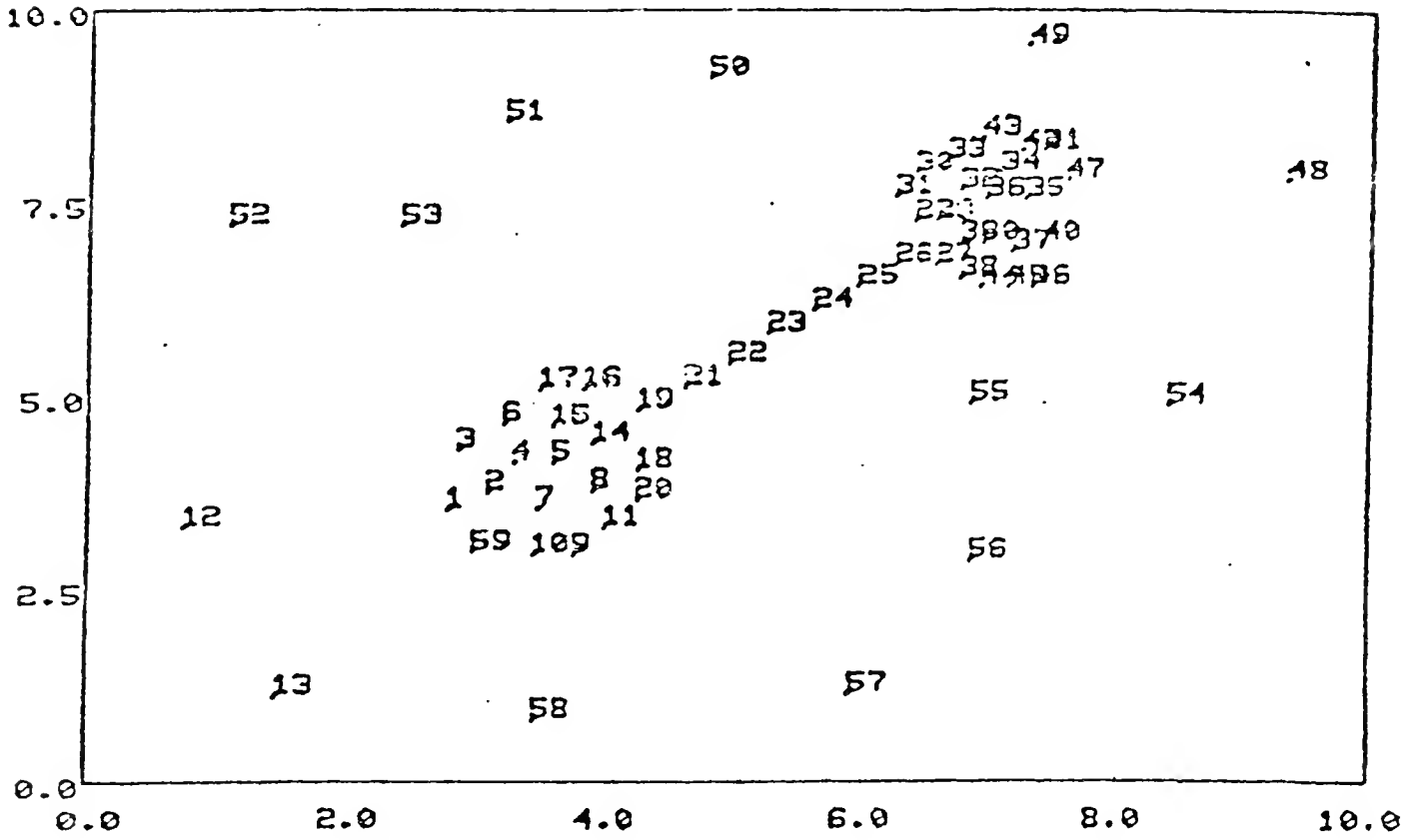


Fig. 3a Scatter-plot of the generated bivariate sample (N = 60) used in Experiment 3. Observation numbers are plotted next to the observations.

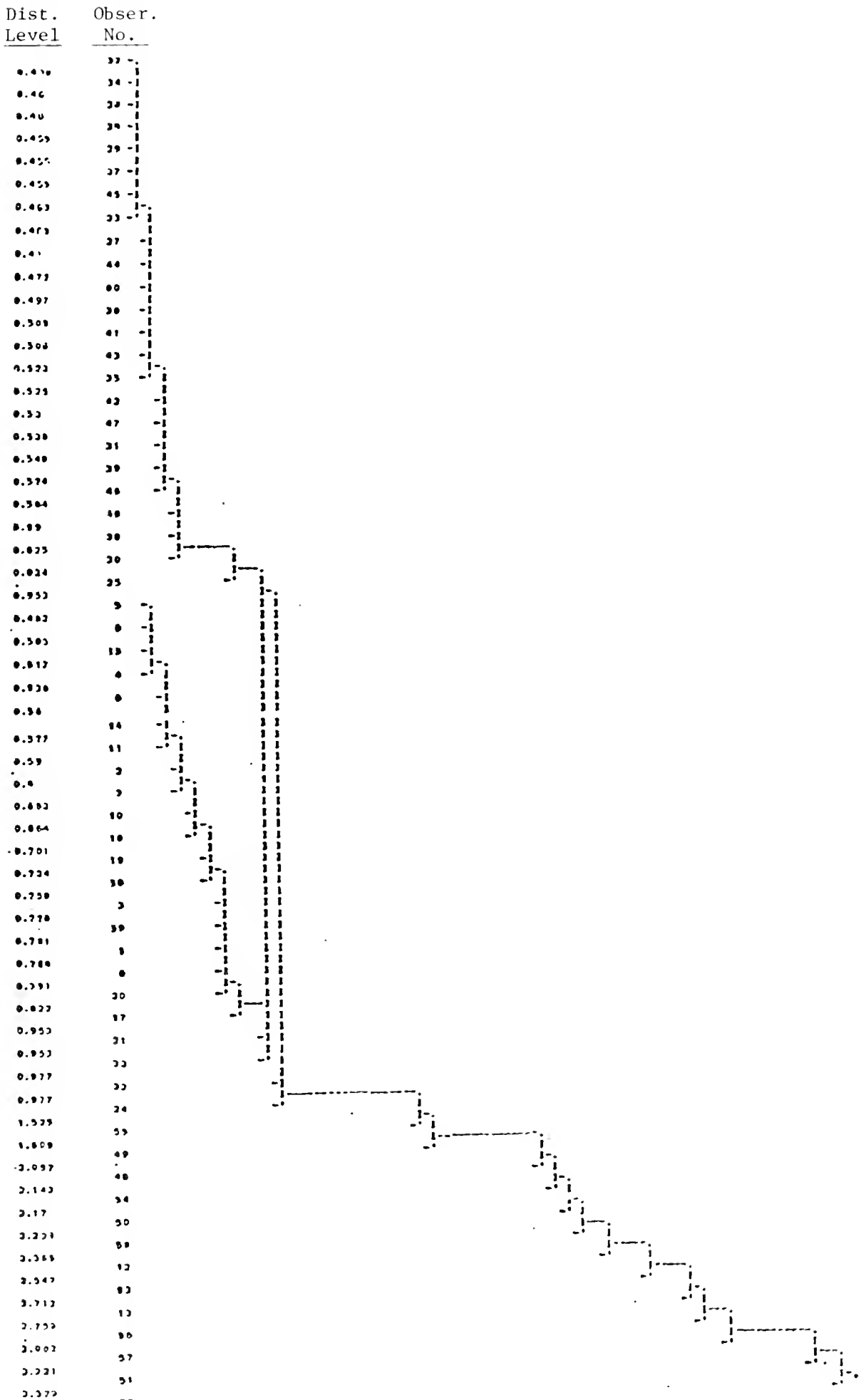


Fig. 3 b Tree of sample high-density clusters for data shown in Fig. 3a, derived from the kth nearest neighbour density estimate using k = 4.

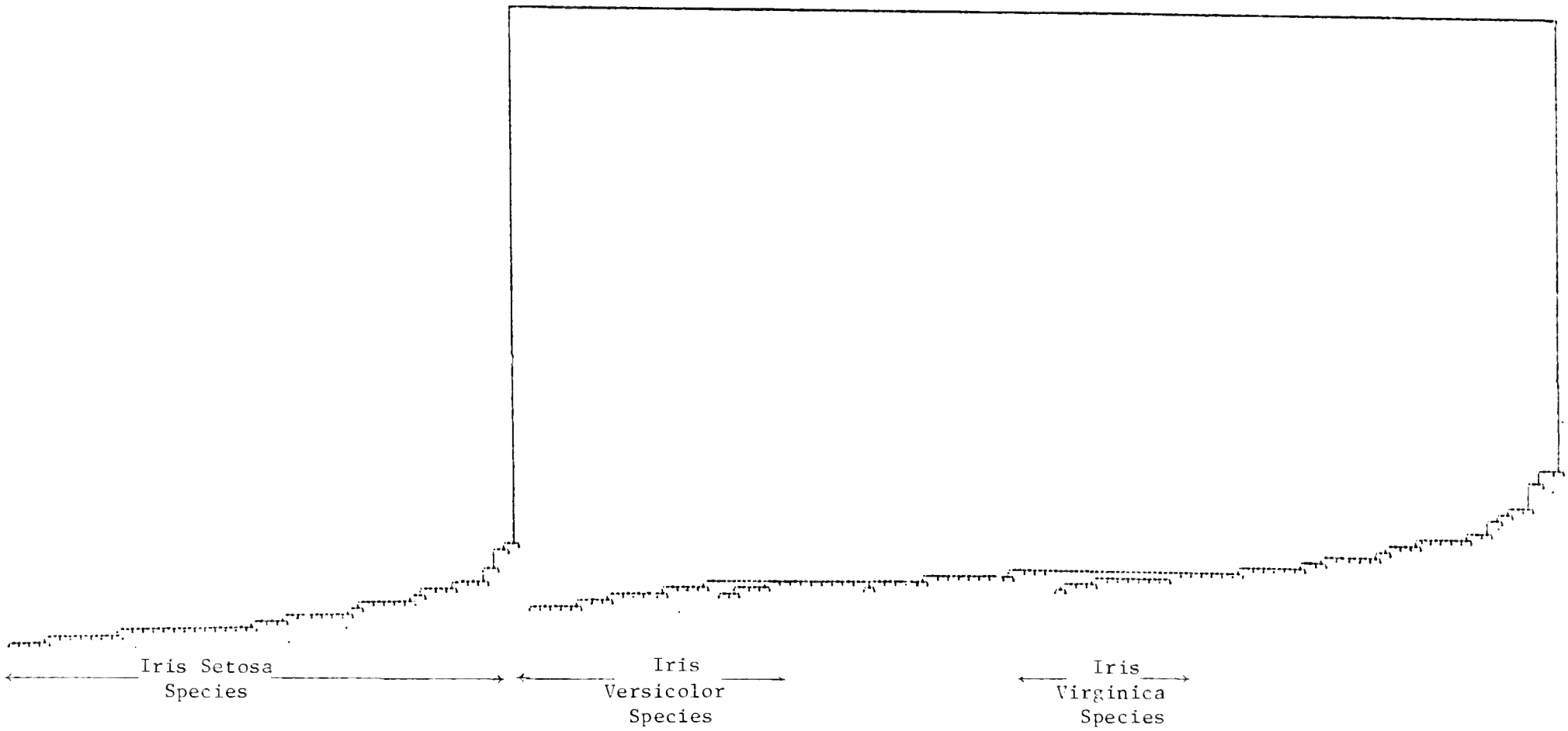


Fig. 4. Hierarchical Clustering obtained for the Iris data using the kth nearest neighbour method (K = 8).

REFERENCES

Anderberg, M.R. (1973). Cluster Analysis for Applications. New York: Academic Press.

Blashfield, R.K., and Aldenderfer, M.S. (1978). "The Literature on Cluster Analysis". Multivariate Behavioral Research, 13, 271-295.

Carmichael, J.W., George, J.A., and Julius, R.S. (1968), "Finding natural clusters". Systematic Zoology, 17, 144-150.

Cormack, R.M. (1971). "A Review of Classification". Journal of the Royal Statistical Society, Series A, 134, 321-367.

Devroye, L.P., and Wagner, T.J. (1977). "The strong uniform consistency of nearest neighbour density estimates". Annals of Statistics, 5, 536-540.

Everitt, B.S. (1974). Cluster Analysis, Halsted Press, New York: John Wiley.

Fisher, R.A. (1936). "Use of multiple measurements in taxonomic problems". Ann. Engen. Lond., 7, 179-188.

Friedman, H.P., and Rubin, J. (1967). "On some invariant criteria for grouping data". Journal of the American Statistical Association, 62, 1159-1178.

Friedman, J.H., Bentley, J.L., and Finkel, R.A. (1975). "An algorithm for finding best matches in logarithmic time." SLAC-PUB-1549, Feb., 1975.

Gnanadesikan, R. (1977). Statistical Data Analysis for Multivariate Observations. New York: John Wiley & Sons, 317-221.

Gower, J.C. and Ross, G.J.S. (1969). "Minimum spanning trees and single linkage cluster analysis". Applied Statistics, 18, 54-64.

Hartigan, J.A. (1975). Clustering Algorithms. New York: John Wiley & Sons.

_____(1977a). "Distributional problems in clustering", in Classification and Clustering. ed. J. Van Ryzin, New York: Academic Press.

_____(1977b). "Clusters as modes", in First International Symposium on Data Analysis and Informatics, Vol. 2, IRIA, Versailles.

_____(1979). "Consistency of single linkage for high-density clusters". Unpublished manuscript, Department of Statistics, Yale University.

_____, and Wong, M.A. (1979). "Hybrid Clustering". Proceedings of the 12th Interface Symposium on Computer Science and Statistics, ed. Jane Gentleman, U. of Waterloo, Press. pp. 137-143.

Moore, D.S. and Yackel, J.W. (1977). "Consistency properties of nearest neighbour density function estimators. Annals of Statistics, 5, 143-154.

Sneath, P.H.A. (1957). "The application of computers to taxonomy", Journal of General Microbiology. 17, 201-226.

_____, and Sokal, R.R. (1973). Numerical Taxonomy. San Francisco: W.H. Freeman.

Sorenson, T. (1948). "A method of estimating groups of equal amplitude in plant sociology based on similarity of species content". K. Danske Vidensk. Selsk. Skr. (Biol.), 5, 1-34.

Spath, H. (1980). Cluster Analysis Algorithms. Halsted Press, New York: John Wiley.

Wishart, D. (1969). "Mode Analysis" in Numerical Taxonomy. edited by A.J. Cole, New York: Academic Press.

Wong, M.A. (1980). "A hybrid clustering method for identifying high-density clusters". Working Paper #2001-80, Sloan School of Management, Massachusetts Institute of Technology.